Alerts "Key Numbers"

The final draft of **DMTN-102: "LSST Alerts: Key Numbers"** can be found here.

The draft RFC is still in progress and can be found here.

Original Project Goal: Produce a document summarizing key numbers for the Alert Production pipeline, e.g. expected numbers of alerts, completeness /contamination, data volumes, etc. This should include descriptions of how these numbers were derived and details on their meaning (i.e. best estimates vs. upper bounds or sizing requirements).

Melissa, Eric, Leanne, Colin, et al.

- **MOTIVATION**: bandwidth for Alerts delivery to multiple Brokers will be limited
- **GOAL**: define and estimate a set of "Key Numbers" related to Alert Distribution
- include the basis information, assumptions, and derivation method for all
- be clear about which are estimates, calculations, boundaries, or limits
- example quantities might include:
    - number per visit of Alerts, new sources, types of sources, false detections
    - Alert packet sizing, full Alert Stream volume, available bandwidth
    - mini-broker deliverables (number of users, filter latency, alerts/visit)
- **EXPECTED OUTCOME**: a public-facing document, target audience broker dev's

2019-06-05: RFC-600 is circulating and refining the proposed LCR.

2019-02-19: DMTN-102 has been released, but the draft RFC proposing changes to the documentation which was inspired by the background work for DMTN-102 remains stalled.

2019-01-25 DM-SST: a trimmed DMTN-102 and a RFC draft.

(1) DMTN-102 has been streamlined, with each topic having a brief key number statement in boldface font followed by an explanation. Where applicable, formal requirements are cited in the right-hand margin (like the DPDD). The text of DMTN-102 has been worked so as not to depend on the acceptance of the proposed RFC (below), but it is likely that future edits to DMTN-102 will be necessary. The one outstanding item to do before DMTN-102 should be considered "finished" is to provide minimum (maximum) alert packet size estimates that represent alerts with no history (11 months of forced photometry) in Section 2.3 (EB is working on it).

(2) The draft RFC to propose changes to the documentation which will clarify requirements regarding the alert stream is currently in the GitHub repo for DMTN-102 (file draft_rfc.txt). It will be posted to JIRA once we agree that it is ready.

2018-12-14 DM-SST: issues and topics for conversation (since resolved).

(1) Reading has turned up some potential tensions/omissions in the formal requirements, for which I propose changes to some documentation. Are they necessary/desired?

- Section 2.1 "Alert Release Timescale": potentially update the OSS and the DMSR to include OTR1=98%, the acceptable fraction of alerts released within a latency of OTT1, which is defined in the LSR but not flowed down.
- Section 2.2 "Number of Alerts per Visit": potentially update the DMSR to include nAlertVisitAvg (or transN) which exist in the LSR and OSS but are not flowed down.
- Section 2.7 "Delayed/Failed Alert Distribution": the OSS's definition of sciVisitAlertDelay could be updated to include OTR1=98%
- Section 2.7 "Delayed/Failed Alert Distribution": the OSS's definitions of sciVisitAlertDelay and sciVisitAlertFailure could be incorporated into the DMSR
- Section 2.7 "Delayed/Failed Alert Distribution": should there be an LSR specification which is like OTR1 (which specifies that = 98% of alerts per visit must not be delayed beyond OTT1) but specifies the acceptable fraction of alerts per visit which fail to be released?

The above are now in a draft RFC.

(2) Several sections have placeholders to discuss contingency plans for when requirements are breached (e.g., more than 40000 alerts/visit); should we discuss such things in this document?

- Section 2.1 "Alert Release Timescale": will delayed alerts distributed with latency >OTT1 be flagged in some way? Is that even possible?
- Section 2.2 "Number of Alerts per Visit": what happens when the number of alerts exceeds 40000? Are they issued with a delay?
- Section 2.7 "Delayed/Failed Alert Distribution": for delayed alerts, LSR-REQ-0025 states that " the remaining transients so detectable must still be identified and recorded at the next processing opportunity", but this is not this flowed down to DMSR and so it is unclear what "the next processing opportunity" means.

The above are not urgent enough for an RFC and beyond the scope of the key numbers document.

(3) Several sections have placeholders in which we could include more scientific motivation for the values of formal requirements (or remove what I've included).

- Section 2.1 "Alert Release Timescale": we could expand on the science drivers for OTT1 = 1 minute, which are briefly mentioned in the SRD.
- Section 2.2 "Number of Alerts per Visit": I've included astrophysical event-rate estimates based on the science book, which we could retain or not.

Decided not to expand on science drivers for OTT1, but did keep the astrophysical event-rate estimates.

(4) For Section 2.3 "Alert Packet Size": Eric is working on more accurate estimates for the alert packet sizing, and options for "lite" alerts.

 This was not immediately resolved, is an outstanding issue.

(5) For Section 2.5 "Number of Selected Brokers": is there a document that we could cite for the allocation of 10 Gbps to the alert stream, which I only find quoted in LDM-612?

 JS reports that no, there is no other document to cite for this.

2018-12-07 DM-SST discussion. MLG took notes and added some answers. All has been incorporated into DMTN-102.

(1) Can anyone think of a quantity that is missing? (And would we like to survey the broker developers to see what they need to know?)

- alert production timescale
- number of alerts per visit
- fraction of visits with delayed/failed alert distribution
- fraction of alerts per visit with delayed distribution
- fraction of false positives per visit
- alert packet size
- alert stream data rate
- number of selected brokers
- alert database volume
- number of new transients per visit
- mini-broker: number of simultaneous users
- mini-broker: number of alerts per visit returned per filter
- mini-broker: alerts database latency

(2) Alert Distribution Timescale: OTT1 = 60 seconds starts at the end of readout, and ends when the alert is: (a) released to the stream or (b) transmitted to the broker?

- LG: the time ends when the alert is available for pick-up?
- EB: matters how quickly brokers are able to **get** the alert so they don't fall behind; perception issue: if OTT1 ends when alerts are available to pull, but pulling takes a bunch of time, then there might be disappointment and the feeling that alert distribution is **not** "within 60s"
- CS: could have a **separate** budget on the network transfer and not include it in OTT1; that second budget is then the one that constrains the number of selected brokers; thus keeping OTT1 as a processing and alert 'delivery' time
- EB: better to keep it outside OTT1 because then we can also support more streams while letting it take longer while also meeting OTT1; it is likely the community would favor more streams even if it means a small (e.g., 60 sec) delay
- LG: for the key number, we should define OTT1 as "available to the stream"

(3) Alert Distribution Science Validation: OTR1 = 98% of alerts per visit must be transmitted within OTT1 (LSR-REQ-0025), but this value does not seem to have been flowed down to the OSS or DMSR, and is not used to define whether a visit has experienced "successful" alert distribution. The OSS requires that <0.1% of visits experience "failed" alert distribution (no alerts) and <1% of visits experience "delayed" alert distribution (not completed by OTT1). It seems the LSR and the OSS are not in agreement on what constitutes a failure of alert distribution. Thoughts?

- it is captured in science verification, and pretty much as-is, so OK to include in the key numbers document like this too

(4) False Positives: I had not realized that 50% of the alerts distributed with transSNR>5 are expected to be false positives. I'm not sure the community realizes this either. This might be a shock? Or are the broker developers cognizant of this?

- needs a discussion by EB CS and MLG (3-13 minutes)

(5) Alert Packet Sizing: Is a stream of "lite" alert packets an option on the table? E.g., no stamps, or no history.

- yes that's going to be an option; integrate into key number document

(6) Alert Stream Data Rate: I've quoted a time-average and a "peak" Mbps if all alerts released in 5 seconds; how the the streaming actually planned to happen?

- still tbd

(7) Number of Selected Brokers: I thought the minimum number of 4 was a formal requirement but could not find this anywhere.

- consensus that there is no formal requirement on the number of brokers served (not part of operations readyness)
- there is a formal requirement on the AFS existing, but we should make sure it gets into the science validation process (within DM)
- JS: there is a section on network load testing for community brokers but it's under EPO (LSE-79 Section 3.5.3)

Preliminary notes, which have since been incorporated into DMTN-102.

- this will be a living document as some values will be refined in operations
- data volumes ("Because of the large data volume of the alert stream (several TB per night)" S.4.2 LDM-612)
- "The mini-broker is required to support 100 simultaneous users" S.2.2.4 LDM-612
- 10 Mb/s from NCSA will go into LDM-612

Notes from Gregory Dubois-Felsmann during DM-SST F2F 11/05/18 via Slack #dm-sst:

Regarding the alerts-per-visit specification:
1) The SRD says that the design spec for "The minimum number of candidate transients per field of view that the system can report in real time." is 10,000 (stretch 100,000; minimum 1,000). It also uses the equivalent language "The system should be capable of reporting such data for at least transN candidate transients per field of view and visit.". (edited)
2) This is flowed down essentially verbatim to LSR-REQ-0101, where the number is still 10,000 and the language is "The minimum number of optical transients for which data can be reported per visit." However, reflecting ambiguities in what people were _saying_ back then, the LSR also says in its non-normative text: "It is unclear whether the SRD specification of transN refers to the number of alerts that can be generated for a single visit (i.e. an instantaneous limit), or the number per visit averaged over time."
3) This is flowed down to OSS-REQ-0193, where the normative text says "The LSST Data Management system shall be sized to accommodate an average value of at least *nAlertVisitAvg* [10,000] alerts generated per standard visit while meeting all its other requirements. Performance shall degrade gracefully beyond that limit." Non-normative text in the "parameter description" for *nAlertVisitAvg* says "Minimum number of alerts required to be accommodated from a single standard visit". (edited)
3a) This language was changed by LCR-145 in 2013. The key wording in the LCR is "Reword alerts-per-visit requirement to indicate that it is an average value that will be exceeded (potentially with increased delivery latency after that point). (Clarification; more stringent requirement on DM; no effect on other subsystem.)".
4) Triggered by this, the science inputs to the sizing model in LSE-81 were changed in internal version v20 of this spreadsheet to show 10,000 as an average requirement and 40,000 as a peak requirement. This change made it to Docushare in Docushare v21, internal version v23, in September 2013.
4a) From what I can tell, that version wasn't ever formally approved by the CCB, but a further change to internal version v24 was CCB-approved under LCR-160 a couple of weeks later, in time for the FDR.
So designing for 10K average and 40K peak is definitely our baseline.